Published in: Proceedings des 9. Hildesheimer Evaluierungs- und Retrievalworkshop (HIER 2015), Universität Hildesheim: Hildesheim, 2015. Pp. 63 - 75.

Scales and Scores

An evaluation of methods to determine the intensity of subjective expressions

Josef Ruppenhofer, Jasper Brandes, Petra C. Steiner

Hildesheim University
Hildesheim, Germany
{ruppenho|brandesj|steinerp}@uni-hildesheim.de

Abstract

In this contribution, we present a survey of several methods that have been applied to the ordering of various types of subjective expressions (e.g. *good* < *great*), in particular adjectives and adverbs. Some of these methods use linguistic regularities that can be observed in large text corpora while others rely on external grounding in metadata, in particular the star ratings associated with product reviews. We discuss why these methods do not work uniformly across all types of expressions. We also present the first application of some of these methods to the intensity ordering of nouns (e.g. *moron* < *dummy*).

1. Introduction

While there is much interest in intensity ordering for application within sentiment analysis, the ability to assess the intensity associated of scalar expressions is a basic capability that NLP systems in general need to have. It is necessary for any NLP task that can be cast as a textual entailment problem, such as IR, Q&A, summarization etc. For instance, as illustrated by de Marneffe et al. (2010), when interpreting dialogue (A: *Was it good?* B: *It was ok / great / excellent*.), a yes/no question involving a gradable predicate may require understanding the entailment relations between that predicate and another contained in the answer.[9]

[9] The intensity ordering task within sentiment analysis can also be understood as an entailment problem, which is prefigured e.g. by Horn's (1976) discussion of conversational implicatures of scalar predicates.

Among gradable linguistic expressions, adjectives are the best-studied class. Various methods have been explored, some of which we will experiment with, namely phrasal patterns (Sheinman 2013; Melo and Bansal, 2013); using star ratings (Rill et al., 2012); extracting knowledge from lexical resources (Gatti and Guerini, 2012); and collostructional analysis (Ruppenhofer et al., 2014).

Less work has gone into the scalar properties of adverbs. Rill et al. (2012b) studied them indirectly in the context of ordering complex adjective phrases containing intensifying adverbs. In submitted work, we have experimented with extending and adapting the methods used for adjectival intensity ordering for use with adverbs.

As far as we know, only work in theoretical linguistics has analyzed intensity orderings among nouns (Morzycki, 2009).

2. Corpora and published ratings

For our experiments we use three corpora. The BNC and ukWaC are used to compute association measures and to mine for linguistic patterns. The Liu corpus of Amazon product reviews is used to project star ratings onto linguistic units. In addition, we evaluate Taboada et al.'s lexical resource as a source of intensity information.

Corpora	Tokens	Reference
Liu	~ 1.06 B	Jindal and Liu,
BNC	~ 0.1 B	Burnard, 2007
ukWaC	~ 2.25 B	Baroni et al., 2009
Lexicon	Entries	Reference
SoCaL	216 intensifying adv., 1549 nouns, 2827 adjectives	Taboada et al., 2011

Table 1. Corpora and published ratings

3. Scales

For the adverb ordering task, we use adjectives from 4 different semantic scales. These are shown in Table 2 together with their classification following Paradis (1997, 2001). The adverbs we used are shown below in Table 3, sorted into the classes defined by Paradis (1997). For the items used in the adjective ordering task, we refer to Ruppenhofer et al. (2014). The items of the noun ordering task are presented in Table 4.

Adjective	Scale	Pol.	Type
dumb	Intelligence	neg	scalar
smart	Intelligence	pos	scalar
brainless	Intelligence	neg	extreme
brainy	Intelligence	pos	extreme
bad	Quality	neg	scalar
good	Quality	pos	scalar
mediocre	Quality	neg	scalar
super	Quality	pos	extreme
cool	Temperature	neg	scalar
warm	Temperature	pos	scalar
frigid	Temperature	neg	extreme
hot	Temperature	pos	extreme
short	Duration	neg	scalar
long	Duration	pos	scalar
brief	Duration	neg	scalar
lengthy	Duration	pos	scalar

Table 2. Classification of adjectives used

Maximizer	Booster
absolutely	awfully
completely	extremely
perfectly	very
quite	highly
Moderator	Diminisher
quite	slightly
fairly	a little
pretty	somewhat
Approximator	Control
almost	no adverb

Table 4. Classification of adverbs used

Intelligence		Expertise	
positive	negative	positive	negative
Einstein, genius, brain, brainiac, superbrain, sage	blockhead, cretin, dimwit, doofus, fathead, fool, half-wit, idiot, imbecile, moron, nitwit	ace, adept, buff, champion, expert, guru, master, maven, pro, specialist, star, superstar, virtuoso, whiz	neophyte, newbie, novice

Table 3. Nouns used

4. Gold Standards

For all the items from the different scales, we elicited ratings using the online survey tool LimeSurvey.[10] We recruited our subjects from Amazon Mechanical Turk (AMT), specifying the following qualifications: US residency, a HIT-approval rate of at least 97%, and 500 prior completed HITs.

[10] www.limesurvey.org

The surveys typically used several parallel surveys, each eliciting data for subsets of our items, to be completed by non-overlapping sets of participants, which we controlled by checking AMT worker IDs. In each survey, participants were first asked for metadata such as age, residency, native language etc. Each survey used pairs of main and distractor block and was concluded at the end by a block in which feedback / comments on the survey was invited. All items were rated individually. The blocks and the items in the blocks were randomized so as to minimize bias. Participants were asked to use a horizontal slider, dragging it in the desired direction, representing polarity, and releasing the mouse at the desired intensity, ranging from −100 to +100.

5. Methods

5.1 Horn patterns

Horn (1976) put forth a set of pattern-based diagnostics for acquiring information about the relative intensity of linguistic items that express different degrees of some underlying property. The complete set of seven diagnostics is shown in Table 5.

For all patterns, the item in the Y slot needs to be stronger than that in the X slot. The two slots can be filled by different types of expressions such as adjectives, nouns, and adjectives, as shown by the following examples.

(1) It's not just *entertaining* but *hilarious*. (adjectives)
(2) Peter's a *dummy*, or even an *idiot*. (nouns)
(3) This is *very* good, if not *extremely* good. (adverbs, with adjective held constant)

Based on the frequencies with which different items of a specific type occur in the X and Y slots, we can induce a ranking of these items.

X (,) and in fact Y	not only X(,) but Y
X (,) or even Y	not X, let alone Y
X (,) if not Y	not Y, not even X
be X (,) but not Y	

Table 2 Horn patterns

5.2 MeanStar

Another corpus-based method we evaluate employs mean star ratings derived from product reviews, as described by Rill et al. (2012b). Note that this method uses no linguistic properties intrinsic in the text. Instead, it derives intensity for items in the review texts from the numeric star ratings that reviewers (manually) assign to products. Generalizing the approach of Rill et al. (2012b) to any kind of simple or complex unit, we define the intensity score for a unit as the weighted mean of the star ratings

$$SR_i = \frac{\sum_{j=1}^{n} \quad S_j^i}{n}$$

where *i* designates a distinct unit, *j* is the *j*-th occurrence of the unit, S^i_j is the star rating of *i* in *j*, and *n* is the total of observed instances of unit *i*.

5.3 Collexeme analysis (Collex)

Collexeme analysis (Gries and Stefanowitsch, 2004) exploits the association strength between linguistic units and the constructions that they can occur in. For instance, in the case of adjectives, one assumes that adjectives with different types of intensities co-occur with different types of adverbial modifiers. End-of-scale modifiers such as *extremely* or *absolutely* target adjectives with a partially or fully closed scale (in the sense of Kennedy and McNally (2005)), such as *brilliant* or *outstanding*, which occupy extreme positions on the intensity scale. 'Normal' degree modifiers such as *very* or *rather* target adjectives with an open scale structure, such as *good* or *decent*, which occupy non-extreme positions.

To determine a linguistic unit's preference for one of two constructions, the Fisher exact test (Pedersen, 1996) is used. It makes no distributional assumptions and does not require a minimum sample size. The direction in which observed values differ from expected ones indicates a preference for one construction over the other and the p-values are taken as a measure of the preference strength.

In the case of adjectives, our hypothesis is that e.g. an adjective A with greater preference for the end-of-scale construction than adjective B has a greater inherent intensity than B.

Note that Collex produces two rankings, one representing the degree of attraction to one of the constructions. To obtain a global intensity ordering, they need to be combined. In the case of ordering adjectives, the positive/negative adjectives being attracted to the extreme modification construc-

tion were put at the top/bottom of the ranking. The set of adjectives that prefer the normal modification construction are placed between the extreme positive and negative sets. Here, the positive/negative adjective least attracted to the normal construction immediately adjoins the positive/negative adjective least attracted to the extreme construction. Adjectives that have no preference for either construction are finally inserted in between the positive and negative adjectives attracted to the normal construction.

For adverbs, we consider the adjective-adverb nexus in the opposite direction: the adverbs are the units to score and classes of adjectives define the different constructions. For nouns, we can proceed in simple analogy to the case of adjectives, except that the modifiers of nouns are adjectives such as *high* or *utter* rather than adverbs such as *highly* or *utterly*.

6. Experiments

6.1 Adjectives

In earlier work (Ruppenhofer et al., 2014), we compared the performance of our methods on both subjective adjectives as well as objectives ones. We found Collex to give good performance for both types of adjectives. While de Melo and Bansal (2013) report very good results using Horn patterns, we prefer the use of Collex because it does not need web-scale data (Google 5-grams), working even on 'smaller' corpora such as the BNC, and is computationally simpler than the sophisticated interpolation approach applied by those authors. The MeanStar method was slightly better than Collex for subjective adjectives but very low-performing for objective ones. Of the lexical resources we considered, SoCAL had the best results. However, SoCAL has coverage gaps for objective adjectives.

6.2 Adverbs

Horn patterns cannot be used for adverbs, at least not currently. In the ukWaC, there are very few instances of Horn's 7 patterns that have two different adverbs but the same adjective in the X and Y slots. The frequency of relevant adverb-adjective instances is in fact significantly lower than that of simple adjective instances. On web-scale data, this approach might still become feasible. However, it is currently not feasible because for the smallest pattern to host two adverb-adjective pairs in the X and Y slots, one would already need 6-grams, whereas only 5-grams are available.

The collostructional approach also did not perform well, counter to our initial hopes and expectations. Using the same ranking strategy that Ruppenhofer et al. (2014) employed for adjectives (cf. section 5.3) but with adjectives and adverbs switching roles, produced very low correlation results below 0.2. In hindsight, we believe that this is due to a significant asymmetry between adverbs and adjectives. Among adjectives, the extreme and scalar subgroups are the largest and they tend to be well separated: scalar adjectives tend not to have intensities as great or greater than extreme adjectives. Adverbs are different. First, the gold standard data suggest that adverbs in distinct classes do not have separate bands of scaling effect. For instance, of all adverbs, *extremely*, a booster, has the highest scaling effect, at least matching if not outdoing maximizers such as *utterly* and *absolutely*. And while moderators and diminishers are separated pretty well in the human ratings, the approximator *almost* is sandwiched among the diminishers. The Collex approach is not set up to handle this constellation well since, as shown in Figure 1, it expects to find maximizers and approximators to be most drawn to limit and extreme adjectives, and boosters, moderators and diminishers to be attracted by scalar adjectives. Thus, maximizers and approximators should have similar and consistently higher scaling effects than the other types of intensifiers. Reality fails to comply with this assumption and accordingly we obtain poor results. Collex is thus a one-way strategy: it can rank adjectives based on adverbs but not the other way around.

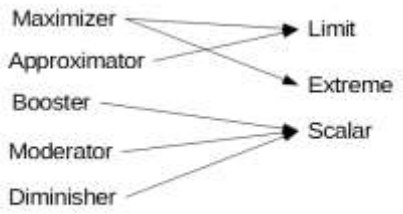

Figure 1. Adverb-adjective interaction

SoCaL's intensifier lexicon has good coverage for our adverbs and provides intensity scores for all of the items. A ranking of our adverbs as obtained by these intensity scores produces near-perfect correlations with our gold standard (0.97). Nonetheless, a drawback of relying on lexical resources for information on the scaling effect of adverbs is that, while there is class of frequently used and highly grammaticized ones, such as the ones we considered here, there is a much larger, fluid set of less grammaticized adverbs (e.g. *preternaturally*) that lexicons will be unlikely to ever fully cover. For these cases having a corpus-based method is necessary.

We finally consider the MeanStar approach. That approach should not in theory be useable directly with adverbs by themselves since they do not have an inherent intensity like adjectives or nouns do but instead act upon the intensity of the predicates they modify. Nevertheless, as a baseline measure we tried the brute force approach of projecting star ratings onto adverbs regardless of the adjectives. The results were better than what we obtained with the collostructional approach: a correlation of 0.283 when using all instances of adverbs found anywhere, and a correlation of 0.446 when only taking into account instances occurring in review titles. This difference between review bodies and titles has been observed before by Rill et al. (2012b) and stems from the fact that titles tend to more straightforwardly match the tenor of the star rating, while review bodies may offer discussions of pros and cons that do not align as cleanly with the star rating given.

Intuitively, if we want to improve upon the adverb-only baseline, we had best taken into account the adjectives being modified by the adverbs. Ideally, we would find every adverb we want to rank used in combination with every adjective that we want to work with. On that basis, we could learn to 'factor' out the effect of the adverb by comparing the scores of adverb-adjective combinations involving the same adjective, to each other and to the score of the unmodified adjective. However, here too, we run up against the actual distribution, which is not as we would like it to be. As the log-log plot in Figure 2 shows, there are many adjectives that occur with a few adverbs and few adjectives that occur with many. We therefore do not find all the combinations that we would need to have so that we could produce per-adjective rankings of the adverbs, which we could then combine into a global ranking of the adverbs.

This distributional fact doomed the first method that we experimented with, which tried to integrate the relative intensity differences between adverb-adjective combinations and other combinations and the simple adjectives, by observing which combinations tend to have greater scores than others. Technically, this was a use of the Borda count method from Voting theory, where voters rank some number of candidates in their order of preference. The adjectives can be thought of as the 'voters' on the ranking of the adverbs. However, this approach performed badly because with our data, we fail to satisfy a core assumption of Borda count, namely that candidates not voted for (i.e. unobserved adverb-adjective combinations) should be ranked lower than any candidates voted for (i.e. observed adverb-adjective combinations).

Actually, even the combinations per adjective that we do find are somewhat deceptive. As shown by the work of Desagulier (2014), adjectives may prefer to co-occur for instance with different moderators depending on the specific word sense involved. As an illustration, consider that in the ukWaC corpus

the combination *pretty cool* is almost 100% associated with the desirability sense of *cool* found in e.g. *cool idea*! By contrast, the combination *fairly cool* is almost exclusively used in the temperature sense found e.g. in *cool weather*. Any corpus-based method must thus make the bet that the most frequent readings of most adjectives will nevertheless belong to the same adjective type in the sense of Paradis and can thus be conflated together.

Accepting that one needs to deal with lemma-level data and pursuing an approach that tries to capture an adverb's scaling effect against the simple adjective, there remains the problem of how to conceive of that scaling effect. In the context of research on review mining, Liu and Seneff (2009) model it as the difference between the intensity of an adverb-adjective combination and the intensity of single adjective. They did not, however, evaluate their model directly against human ratings as a gold standard but only extrinsically as part of an automatic system. It is therefore not clear how well their model of adverb intensity works.

In work of our own that is currently under review, we have pursued a different approach of conceiving of the scaling effect. Basically, we try to capture the relative scaling effect rather than the absolute distance. For example, if we measure the difference between *absolutely good* and simple *good*, and between *absolutely perfect* and simple *perfect*, then on the Liu and Seneff approach, *absolutely* will seem to have a weaker effect on the adjective *perfect* than on the adjective *good* because *perfect* has a higher intensity to start with. Our approach instead asks: how far does the adverb move the adjective's intensity towards the end of the scale, relative to the available distance to be covered? On that approach, the scaling effect of *absolutely* will seem substantial even when applied to *perfect*. We obtained very good results for this method but comparing it to the Liu and Seneff approach remains for future work.

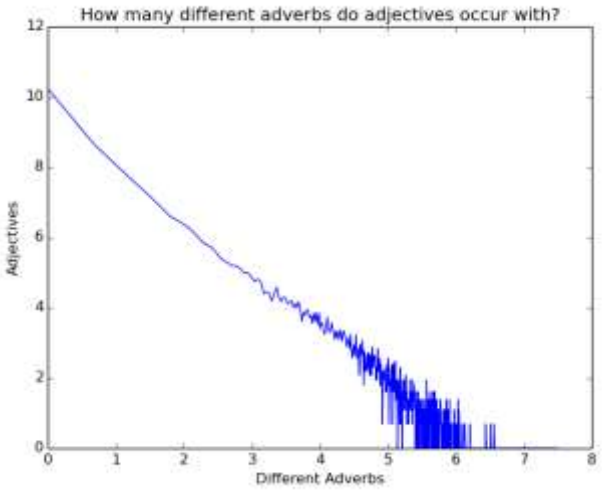

Figure 2. Adverb-adjective co-occurrence in the ukWaC

6.3 Nouns

As for the adjectives and adverbs, for the nouns we also come across severe coverage issues for Horn's patterns. For none of the 7 patterns, we find instances where there are two different nouns from one of the two examined scales. This even holds true if we loosen the constraint and allow up to three additional tokens in between the determiner and the noun. Thus, we currently see no way of using Horn's patterns for nouns.

SoCaL's coverage for nouns is poorest across the three types of expressions investigated in this study: there are intensity scores for only 4 of the 17 intelligence nouns and only for 5 of the 17 expertise nouns. As such, at least for our two scales, SoCaL cannot be used for the intensity ordering of nouns referring to the same scale.

MeanStar produces low positive correlations (0.2) for the intelligence nouns and medium positive correlations (0.51) for the expertise nouns when performed on the review titles. While these results are not good, for the intelligence nouns they can be attributed to the low frequencies of these nouns in the review titles. Coverage, however, is quite good with 15 intelligence and 16 expertise nouns occurring in the review titles.

Finally, we report on the results for Collex. We followed the same approach as for the adjectives, only that for the nouns, we replaced the adverbs with adjectives (i.e. *high* instead of *highly* and *utter* instead of *utterly*). We distin-

guish between two constructions a noun can occur in: modification by 'end-of-scale' adjectives such as *utter* or *complete* or by 'normal' adjectives such as *big* or *slight*. We assume that nouns which are more attracted to 'end-of-scale' adjectives have a higher inherent intensity than nouns that are more attracted to 'normal' adjectives. The ranking approach put forward in Ruppenhofer et al. (2014), yields medium correlations (0.58) for the expertise nouns and high correlations (0.89) for the intelligence nouns.

7. Conclusion

In this paper, we presented a discussion of methods for determining the intensity of subjective expressions. We focused on different semantic scales of English adverbs, adjectives, and nouns. In the case of adjectives and nouns, we have examined both subjective and objective scales.

None of the presented methods works universally well for all considered types of expressions. While Horn's patterns (e.g. 'X or even Y'), one of the two linguistically grounded methods, seem promising, severe coverage issues make this approach currently unusable. The other linguistically motivated method, Collex, works very well for adjectives and quite well for nouns, while for adverbs correlations with a human gold standard are very low. MeanStar produces good correlations for the adjectives and low to medium correlations for adverbs and nouns. Note that the MeanStar approach is dependent on (manually assigned) metadata from a large review corpus which may not be available for all languages. This is also the case for lexical resources which assign intensity ratings to lexical items. SoCaL, a much-cited subjectivity lexicon, fares well for the adjectives and adverbs but has very low coverage for nouns. The pursuit of corpus-based methods is thus necessary for reasons of coverage. Potentially, it is also interesting for building customized intensity ratings, for instance, for the American versus the British variety of English, but also for specific application contexts such as product review mining.

8. References/Literaturverzeichnis

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetti. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. Language Resources and Evaluation, 43(3):209-226.

Burnard, Lou. (2007). Reference Guide for the British National Corpus, Research Technologies Service at Oxford University Computing Services, Oxford, UK.

Desagulier, Guillaume. (2014). Corpus Methods for Semantics, chapter Visualizing distances in a set of near-synonyms, pages 145-178. John Benjamins Publishing Company, Amsterdam, Philadelphia.

de Marneffe, Marie-Catherine, Christopher D. Manning, and Christopher Potts. (2010). Was it good? It was provocative. Learning the meaning of scalar adjectives. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. ACL '10, pages 167-176, Stroudsberg, PA, USA. Association for Computational Linguistics.

de Melo, Gerard G. and Mohit Bansal. (2013). Good, Great, Excellent: Global Inference of Semantic Intensities. In: Transactions of the Association for Computational Linguistics 1:279-290

Gatti, Lorenzo & Marco Guerini (2012). Assessing Sentiment Strength in Words Prior Polarities. Proceedings of the International Conference on Computational Linguistics (COLING). 2012, 361-370.

Gries, Stefan Th. and Anatol Stefanowitsch. (2004). Extending collostructional analysis: a corpus-based perspective on alternations. International Journal of Corpus Linguistics, 9(1):97-129.

Horn, Laurence Robert. (1976). On the Semantic Properties of Logical Operators in English. Indiana University Linguistics Club.

Jindal, Nitin and Bing Liu. (2008). Opinion Spam and Analysis. Proceedings of the international conference on Web search and web data mining (WSDM), pages 219-230.

Kennedy, Christopher and Louise McNally. (2005). Scale Structure, Degree Modification, and the Semantics of Gradable Predicates. Language, 81(2):345-338.

Liu, Jingjing and Stephanie Seneff. (2009). Review Sentiment Scoring via a Parse-and-Paraphrase Paradigm. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 161-169

Morzycki, Marcin. (2009). Degree modification of gradable nouns: size adjectives and adnominal degree morphemes. Natural Language Semantics 17(2):175-203.

Paradis, Carita. (1997). Degree modifiers of adjectives in spoken British English. Lund University Press.

Paradis, Carita. (2001). Adjectives and boundedness. Cognitive Linguistics, (12):47-65.

Pedersen, Ted. (1996). Fishing for Exactness. Proceedings of the South-Central SAS Users Group Conference.

Rill, Sven, Sven Adolph, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schütz, Florian Wogenstein, Roberto V. Zicari, and Nikolaos Korfiatis. (2012a). A phrase-based opinion list for the German language. Proceedings of KONVENS, 305-313.

Rill, Sven, Johannes Drescher, Dirk Reinel, Jörg Scheidt, Oliver Schuetz, Florian Wogenstein, and Daniel Simon. (2012b). A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. Proceedings of the KDD-Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM).

Ruppenhofer, Josef, Michael Wiegand, and Jasper Brandes. (2014). Comparing methods for deriving intensity scores for adjectives. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers, pages 117-122, Gothenburg, Sweden, April. Association for Computational Linguistics.

Sheinman, Vera & Takenobu Tokunaga. (2009). AdjScales: Differentiating between Similar Adjectives for Language Learners. CSEDU 1:229-235.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. (2011). Lexicon-Based Methods for Sentiment Analysis. In: Computational Linguistics 37 (2):267-307.